EDA
Empirische Forschungsmethoden und deskriptive Statistik
- eine Einführung -

TECHNISCHE UNIVERSITÄT DARMSTADT

Studienkolleg
Academic Bridging Courses
für ausländische Studierende

Hans Göttmann

Skript zur Lehrveranstaltung im 1. Semester der G-Kurse

Dieses Skript ersetzt nicht die Teilnahme an der Lehrveranstaltung.
Es erhebt keinen Anspruch auf Fehlerfreiheit und wird regelmäßig aktualisiert.
Die noch fehlerhafte Nummerierung der Abbildungen bitte ich zu entschuldigen.
Hinweise auf Fehler oder Vorschläge für Erweiterungen sind stets willkommen.

Bearbeitungsstand: 1.1.2016

Inhalt

1 Lehrziele der Veranstaltung

In den meisten Sozialwissenschaften spielen empirische Forschungsmethoden eine wichtige Rolle. Im Studium steht deshalb oft das Fach Statistik am Anfang, ein Fach, das für viele Studierende mit Angst und Unsicherheit verbunden ist. Diese Veranstaltung soll Ihnen die Grundlagen liefern, den Anforderungen im Fach Statistik ohne Angst und vielleicht sogar mit Neugier und Interesse zu begegnen. In kreativen Fachgebieten wie Design oder Musik aber auch in Literaturwissenschaft oder Geschichte gibt es Studiengänge ohne Statistik. Aber fast überall sonst müssen Sie sich mit Statistik beschäftigen, auch in Sportwissenschaft oder Pädagogik. In diesem Kurs sollen Sie

- die Grundlagen der quantitativen empirischen Forschung kennenlernen,
- Wortschatz, Methoden und Kenngrößen der deskriptiven Statistik kennen und anwenden.

2 Von der Forschungsfrage zum Fragebogen

Ausgangspunkt einer empirischen Untersuchung ist immer eine möglichst präzise Fragestellung: Was wollen wir wissen bzw. erforschen? Eine mögliche Fragestellung wäre z.B.:

Was ist für deutsche Männer und Frauen eine ideale Familie?

Um diese Frage zu beantworten, könnte man z.B. möglichst viele Leute einfach erzählen lassen, was ihnen zu dem Thema einfällt oder man könnte die Fachliteratur und Tageszeitungen nach Artikeln zu diesem Thema durchsuchen. Beide Methoden sind in den Sozialwissenschaften weit verbreitet, haben aber folgende Nachteile:

- Die Ergebnisse werden von verschiedenen Lesern verschieden interpretiert.
- Die Ergebnisse kann man nur schwer miteinander vergleichen.
- Die Ergebnisse sind nicht eindeutig.
- Man kann die Untersuchung schwer oder gar nicht unter gleichen Bedingungen wiederholen.

Um diese Nachteile zu vermeiden, entwickelt man aus der ursprünglichen Forschungsfrage mehrere Fragen, deren Antworten man genauer einordnen und vergleichen kann. Das nennt man *operationalisieren*. Beim Operationalisieren werden folgende Voraussetzungen geschaffen:

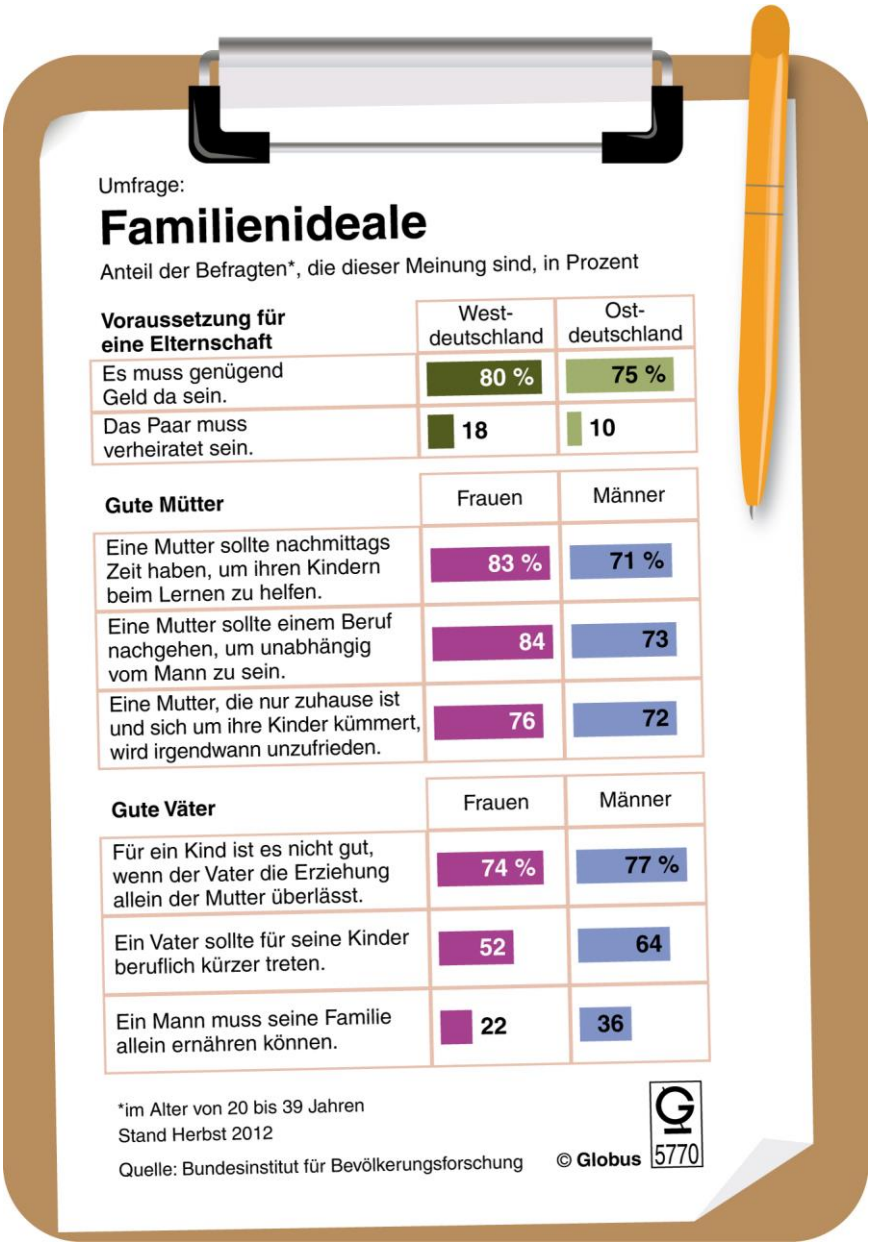
- Die Fragen sind verständlich: Jeder Befragte kann sie verstehen.
- Die Fragen sind präzise: Alle Befragten verstehen die Frage mehr oder weniger gleich.

In einem weiteren Schritt überlegt man nun, wie man die Antworten statistisch auswerten kann. Dazu müssen die Fragen weitere Voraussetzungen erfüllen:

- Die Fragen sind geordnet: Die Reihenfolge der Fragen ist für alle Befragten gleich.
- Die Antwortmöglichleiten sind endlich: Es gibt nur eine begrenzte Menge von Alternativen
- Die Antwortmöglichkeiten sind skaliert: Man kann die Antworten einem Wert auf einer Skala zuordnen.

Diesen Schritt nennt man *standardisieren*.

Das Ergebnis ist ein *Fragebogen*, der die operationalisierten und standardisierten Fragen enthält. Die mit Hilfe des Fragebogens *erhobenen* Daten werden dann statistisch ausgewertet und anschaulich dargestellt, z.B. in Form des folgenden Diagramms.



Umfrage:

Familienideale

Anteil der Befragten*, die dieser Meinung sind, in Prozent

Voraussetzung für eine Elternschaft	Westdeutschland	Ostdeutschland
Es muss genügend Geld da sein.	80 %	75 %
Das Paar muss verheiratet sein.	18	10

Gute Mütter	Frauen	Männer
Eine Mutter sollte nachmittags Zeit haben, um ihren Kindern beim Lernen zu helfen.	83 %	71 %
Eine Mutter sollte einem Beruf nachgehen, um unabhängig vom Mann zu sein.	84	73
Eine Mutter, die nur zuhause ist und sich um ihre Kinder kümmert, wird irgendwann unzufrieden.	76	72

Gute Väter	Frauen	Männer
Für ein Kind ist es nicht gut, wenn der Vater die Erziehung allein der Mutter überlässt.	74 %	77 %
Ein Vater sollte für seine Kinder beruflich kürzer treten.	52	64
Ein Mann muss seine Familie allein ernähren können.	22	36

*im Alter von 20 bis 39 Jahren
Stand Herbst 2012

Quelle: Bundesinstitut für Bevölkerungsforschung © Globus 5770

***Übungsaufgaben**:*

1. *Wie kann man sich den Fragebogen vorstellen, der dieser Grafik zugrunde liegt? Was wurde gefragt? Wie wurde gefragt? Wer wurde befragt? Welche Informationen fehlen?*

2. *Beurteilen Sie die folgenden Fragen: Sind sie operationalisiert? Sind sie standardisiert?*

a) *Wie geht es Ihnen?*
b) *Wie ist Ihre Meinung zu Deutschland?*
c) *Haben Sie schon einmal einen Archäopterix gesehen?*
d) *Welche Noten hatten in Mathematik?*
e) *Welche Note hatten Sie in der letzten Mathematikprüfung?*
f) *Wie gut waren Sie in der Schule in Mathematik?*

3. *Führen Sie im Kurs eine kleine Befragung durch. Erfragen Sie z.B Geburtsdatum, Körpergröße, Anzahl der Geschwister, Schuhgröße usw. Wenn es Spaß macht, können Sie auch noch andere Daten ermitteln.*

3 Was sind Daten?

In der empirischen Forschung interessiert man sich für *Merkmale* (Attribute) in einer Gruppe von Untersuchungseinheiten (z.B. Personen, Länder usw.). Die Gruppe aller Untersuchungseinheiten, in denen Merkmale von Interesse sind, nennt man *Grundgesamtheit* oder *Population*. Wenn wir z.B. die Studierenden unseres Kurses als Population definieren, dann können wir mehrere Merkmale untersuchen, die alle Mitglieder der Gruppe besitzen, wie z.B. Geschlecht, Alter, Größe, Deutschnoten usw.

Es ist klar, dass nicht jeder Student am gleichen Tag geboren ist. Das Merkmal „Geburtsdatum" kann also verschiedene *Ausprägungen* annehmen – es ist veränderlich. Merkmale, die verschiedene Ausprägungen annehmen können, nennt man *Variable*.

Es ist ebenfalls einleuchtend, dass die möglichen Werte der Variable „Alter in Jahren" nicht unbegrenzt sind. Weder gibt es negative Werte aber auch 150-jährige Studierende sind nicht möglich. Die Menge der Werte, die man für eine Variable erwarten kann, heißt *Wertebereich.*

Wenn wir nun die einzelnen Studenten nach Ihrem Geburtsdatum fragen, erhalten wir die verschiedenen Ausprägungen dieser Variable, die *Werte*. An diesem Beispiel wird auch klar, welche Zuordnung zwischen *Merkmalsträger* (Student) und *Merkmalsausprägung* (Variablenwert) gilt: Jeder Student hat nur ein Geburtsdatum, aber theoretisch können mehrere Studenten am gleichen Tag geboren sein. Wir können also jedem Studenten einen Wert zuordnen. Das nennt man *messen*. Beim Messen ordnet man jedem Element der Grundgesamtheit einen Variablenwert zu.

4 Klassifizierung von Merkmalen

Am Beginn der statistischen Auswertung steht die Messung von Merkmalen.

Merkmale kann man unterscheiden

- nach Art der Ausprägungen: qualitativ, quantitativ
- nach der Zahl der möglichen Ausprägungen: diskret, quasi-stetig, stetig
- nach der Struktur des Wertebereiches: nominal, ordinal, Intervallskala, Verhältnisskala

Qualitative Merkmale wie können nicht in Zahlen ausgedrückt werden, sondern nur in beschreibenden Begriffen. Auf die Frage nach Ihrer Lieblingsfarbe werden Sie wahrscheinlich nicht mit einer Zahl antworten.

Quantitative Merkmale kann man messen und mit Zahlen verbinden. Auf die Frage nach der Anzahl der Arbeitslosen in Deutschland erhalten wir eine einfache Zahl, auf die Frage nach dem Körpergewicht eine Angabe aus Zahl und Einheit, z.B. 67 kg.

Ein diskretes Merkmal kann nur endlich viele (abzählbar viele) Ausprägungen annehmen.
Beispiel: Die Anzahl der Studierenden im Kurs kann nur den Wert ganzer Zahlen annehmen.

Ein stetiges Merkmal kann (im Prinzip) alle Werte in einem Intervall annehmen. (überabzählbar viele)
Beispiel: Die Körpergröße der Studierenden im Kurs kann man theoretisch unendlich genau messen.

In der Statistik werden oft Merkmale, die eigentlich diskret sind, als stetig behandelt, um damit besser rechnen zu können. Wenn z.B. der Kurs US-Dollar/Euro 1:0,99090 beträgt, werden Sie dennoch bei der Bank keine 0,99099 Euro bekommen, sondern eben nur 99 Cent.

Durch die Klassenbildung werden stetige Daten immer diskretisiert. (dazu später!)

Merkmale können verschieden skaliert sein. Die Skalierung bestimmt, welche Transformationen (=Rechenoperationen) man mit den Werten durchführen kann.

Nominalskala - Ein Merkmal heißt nominalskaliert, wenn die Ausprägungen Namen oder Kategorien sind, die keine natürliche Ordnung haben, z.B. Geschlecht (männlich, weiblich); Farbe (blau, gelb, grün, rot usw.)

Reihenfolge beliebig;
qualitatives Merkmal; diskret
Rechenoperationen: = und ≠
Aussagen: gleich oder ungleich; Modalwert, Häufigkeitsverteilung

Ordinalskala - Ein Merkmal heißt ordinalskaliert, wenn sich die Ausprägungen ordnen lassen, z.B. Herkunftsregion (Europa, Asien, Afrika, Amerika); Schulnoten (1-6)

Abstand zwischen den Werten nicht definiert; kein definierter Nullpunkt
Rechenoperationen: < oder >
Aussagen: zusätzlich: kleiner als, größer als…; Modalwert, Median

Intervallskala – Ein Merkmal heißt intervallskaliert, wenn die Abstände der Merkmalsausprägungen sinnvoll interpretiert werden können, z.B. Temperatur in Grad Celsius; Datum (TT.MM.JJJJ)
Abstand zwischen den Werten definiert
kein Nullpunkt oder willkürlich festgelegter Nullpunkt
Rechenoperationen: + oder – (Addition oder Subtraktion)
Aussagen: zusätzlich: um wieviel größer bzw. kleiner als…; Modalwert, Median, Quantile

Verhältnisskala (Ratioskala) – Ein Merkmal heißt verhältnisskaliert, wenn es intervallskaliert ist und zusätzlich ein sinnvoll interpretierbarer Nullpunkt existiert, z.B. Körpergröße (in cm); monatliche Ausgaben (in Euro)
Nullpunkt definiert
Rechenoperationen: · oder : (Multiplikation oder Division)
Aussagen: zusätzlich: Verhältnis zwischen den Werten; Mittelwert, Varianz, Standardabweichung

Qualitative Merkmale sind entweder nominal- oder ordinalskaliert. Quantitative Merkmale sind entweder intervall- oder verhältnisskaliert.

Hinweis: In der Praxis nimmt man es mit den erlaubten Rechenoperationen oft nicht so genau und „tut so“, als wären die Daten verhältnisskaliert. So wird z.B. oft eine Durchschnittsnote berechnet, obwohl Schulnoten nur ordinal skaliert sind. Infolgedessen wäre eigentlich keine Division und damit auch keine Berechnung des Mittelwerts möglich.

Übung:

Bestimmen Sie die Merkmalsart (qualitativ / quantitativ? stetig/diskret?) und das Skalenniveau der folgenden Variablen. Begründen Sie jeweils Ihre Entscheidung!

1. *Die Nummern auf den Trikots einer Fußballmannschaft*
2. *Die Tabellenplätze der Bundesliga*
3. *Die Temperatur in der Einheit Kelvin*
4. *Die Geschwindigkeit von 100m-Läufern*
5. *Das Lebensalter der Menschen in einem Land*
6. *Telefonnummern*
7. *Jahreszahlen*
8. *Lieblingsfarbe*
9. *Wahlentscheidung für eine Partei bei der Bundestagswahl*
10. *Benzinpreis pro Liter*
11. *Zufriedenheit mit der IT-Ausstattung in der TU Darmstadt (sehr gut – gut – mittel – schlecht – sehr schlecht)*

5 Die Stichprobe

Wenn die Daten von allen Objekten der Grundgesamtheit erhoben werden, dann spricht man von einer *Vollerhebung*. Das ist z.B. der Fall, wenn wir alle Studierenden des Studienkollegs nach ihrem Alter fragen.

Stellen Sie sich vor, Sie sollen für ein Meinungsforschungsinstitut die Meinung der deutschen Bevölkerung über 15 Jahre zu einem bestimmten Thema herausfinden. Es ist klar, dass Sie in diesem Fall nicht alle Menschen anrufen und befragen können. Das würde bei ca. 70 Millionen Menschen nämlich zu lange dauern. In diesem Fall müssen Sie die Untersuchung auf *eine Teilmenge der Population* beschränken. Eine solche Teilmenge heißt *Stichprobe* (engl: *sample*).

Eine gute Stichprobe soll die gesamte Population repräsentieren. Sie soll *repräsentativ* sein. D.h. alle Aussagen über die Stichprobe sollen so weit wie möglich auch für die gesamte Population gelten. Damit das der Fall ist, müssen (unter anderem) zwei Bedingungen erfüllt sein:

1. Die Stichprobe muss so groß sein, dass alle Merkmalsausprägungen eine Chance haben, in der Stichprobe vorzukommen. Die Größe der Stichprobe ist von der gewünschten Genauigkeit der Ergebnisse abhängig. Dabei gilt: Je größer die Stichprobe, desto genauer repräsentiert sie die gesamte Population.

   Beispiel: Wir möchten untersuchen, wie viel Zeit die Studierenden für ihr Studium aufwenden. Wenn wir dazu nur fünf Leute befragen, ist die Chance groß, dass wir zufällig gerade fünf besonders faule oder fünf besonders fleißige Studenten ausgewählt haben. Die Befragung ist daher nicht repräsentativ.

2. Die Auswahl der Stichprobe muss so sein, dass jedes Element der Population die gleiche Chance hat, in der Stichprobe vorzukommen.

   Beispiel: Wir befragen im obigen Beispiel nur die Studierenden des ersten Semesters. Damit haben die Studierenden der höheren Semester oder die Master-Studenten keine Chance, in der Befragung vorzukommen. Die Befragung ist somit nicht repräsentativ.

Die einfachste Methode der Stichprobenauswahl ist die *Zufallsstichprobe*. Angenommen, die Elemente der Population sind nummeriert und wir brauchen eine Stichprobe der Größe N. Dann lassen wir den Computer eine Reihe von N Zufallszahlen erzeugen und untersuchen die Elemente mit den entsprechenden Nummern. Wenn wir die Bürger einer Stadt befragen möchten, können wir das Adressbuch der Stadt benutzen und dort z.B. jeden zwanzigsten Bürger befragen.

Daneben gibt es noch die *gewichtete Stichprobe*. In vielen Fällen ist nämlich die Verteilung bestimmter Merkmale in der Population schon vorher bekannt. Wir wissen z.B. wie viel Prozent der

Studierenden im G-Bereich und wie viele im T-Bereich studieren. Wir wissen ebenso den Anteil von Frauen und Männern. Mit diesen Information können wir eine Stichprobe zusammenstellen, die dieselben Proportionen wie die Gesamtpopulation.

Die Übertragung der Ergebnisse der Stichprobe auf die Population nennt man auch *Hochrechnung* – ein Wort, das Sie bei Wahlen in Deutschland sehr oft hören und lesen werden.

In wissenschaftlichen Untersuchungen sind nur diese Arten von Stichproben erlaubt. In anderen Bereichen, besonders zu Werbezwecken, werden jedoch oft *Konvenienzstichproben*[1] oder *willkürliche Stichproben* verwendet. Man befragt die Personen, an die man am einfachsten herankommt. So werden z.B. viele Aussagen über die Nutzung von Internet-Angeboten durch Umfragen im Internet gewonnen. Damit fallen alle Menschen aus der Stichprobe heraus, die keinen Internetzugang haben.

Die Ergebnisse solcher Untersuchungen sind somit nicht repräsentativ.

Aufgaben:

1. Ausgangslage: Sie möchten die Studienfachwünsche der Studierenden am Kolleg untersuchen. Schätzen Sie ein, welche Auswahl ausreichend repräsentativ ist. Begründen Sie Ihre Aussage und überlegen Sie, was man ggf. ändern müsste!

- *Ich befrage drei Studierende.*
- *Ich stelle mich vor den Haupteingang und befrage die ersten 40 Studierenden, die hineingehen.*
- *Da ich Angst vor Frauen habe, frage ich lieber nur männliche Studierende.*
- *Ich befrage einfach meinen eigenen Kurs.*
- *Ich befrage alle, deren Vorname mit A beginnt.*
- *Ich befrage alle, die im ersten Semester sind.*

2. Für eine Untersuchung zum Gesundheitszustand der Deutschen werden Patienten im Wartezimmer von Ärzten befragt. Ist das Ergebnis repräsentativ?

6 Die Datenmatrix

Wenn nur eine einzige Variable untersucht wird, erhält man eine einfache Zuordnung zweier Mengen: Der Menge der Merkmalsträger (z.B. Studierende) und der Menge der Ausprägungen einer Variable (z.B. Deutschnote).

Wenn wir uns z.B. für Klausurnoten interessieren, liegt es nahe, nicht nur eine einzige Variable zu betrachten. Schließlich schreiben die Studenten in mehreren Fächern Klausuren. In diesem Fall werden jeweils einer Untersuchungseinheit (einem Studenten) *m* Variable zugeordnet. Da es aber nicht nur einen, sondern *n* Studenten gibt, ergibt sich eine Zuordnung *m*:*n* wobei *n* die Menge der untersuchten Studenten und *m* die Menge der Variablen angibt.

Solche Zuordnungen stellt man sinnvoll in einer *Matrix* dar. Eine Matrix (Pl: *Matrizen*) ist ein Zahlenschema aus *n* Zeilen und *m* Spalten. Die Zeilenanzahl n entspricht der Anzahl der Objekte. Man spricht auch von einer *n* x *m*-Matrix (lies: "n mal m Matrix" oder "n kreuz m Matrix"). Eine 100x20-Messwertmatrix enthält also für 100 Merkmalsträger die Ausprägungen von 20 Variablen.

Aufgabe:

1 Konvenienz: eigentlich „Übereinkunft“; hier eher im Sinn von „Bequemlichkeit“, „Anwendbarkeit“; vgl. engl: convenient

*Sehen Sie sich die Matrix im Anhang (*Anhang 1: Übungstabelle*) an und beschreiben Sie:*

1. *Wie viele Objekte gibt es?*
2. *Welche Merkmale (Variablen) gibt es?*
3. *Welches Skalenniveau haben die Variablen jeweils?*
4. *Welche Ausprägungen (Werte) sind theoretisch möglich?*

7 Messwertklassen (Kategorien)

Bei Variablen mit einer sehr großen Anzahl von Werten werden Messwerte zu *Messwertklassen k* zusammengefasst. Messwertklassen nennt man auch *Kategorien* oder *Intervalle*. Man spricht dann von *gruppierten Daten*. Damit beim Zusammenfassen von Einzelwerten das Skalenniveau nicht verfälscht wird, muss man jedoch dabei einige Regeln beachten:

1. Kategorien von *Nominalskalen* können beliebig zusammengefasst werden, sollten aber inhaltlich begründet werden.

Beispiel: Gruppierung der Familiennamen in einer Behörde - k_1: Anfangsbuchstaben A-E, k_2: Anfangsbuchstaben F-J usw… Begründung: Erleichterung beim Suchen nach Namen.

2. *Ordinalskalen* können durch Zusammenlegen benachbarter Ränge vereinfacht werden.
Beispiel: Um die Anzahl von Fehlern in einer Klausur besser den Noten zuordnen zu können, wäre z.B. folgende Kategorisierung sinnvoll:

Fehlerzahl	0	1-3	4-6	7-9	10-12	>12
Fehlerklasse	1	2	3	4	5	6

Da man theoretisch unendlich viele Fehler machen kann, ist am rechten Rand der Skala eine nach oben offene Randklasse (>12) definiert. Praktisch heißt das: Auch wer 100 Fehler macht, bekommt die Note 6.

3. Bei der Klassenbildung in Intervallskalen sollen die *Klassenbreiten* (kb) der Klassen gleich sein. Die neue Klasse wird durch die Klassenmitte bezeichnet, das ist das arithmetische Mittel der unteren und der oberen Intervallgrenze.

Beispiel:

Kaufpreis in €	1,00	2,00	3,00	4,00	5,00	10,00
Preisklassen	2,00			7,00		

Beachten Sie: Der Wert der neuen Klasse ist nicht das Mittel aus allen Werten innerhalb der Klasse, sondern nur der Intervallgrenzen! Der Wert 7,00 ergibt sich aus (4 + 10) : 2. Das Mittel aller Werte innerhalb der Klasse wäre: (4 + 5 + 10) : 3 = 6,33…

4. Klassen dürfen sich nicht überschneiden.

Beispiel: Die Klasseneinteilung k_1:Deutschnote 1-3; k_2 Deutschnote 3-6 ist unzulässig.

Aufgaben:

1. *In der Übungstabelle im Anhang sind die Werte der Variablen „Alter“ und „Ausgaben“ dargestellt. Bilden Sie für beide Variablen sinnvolle Klassen. Begründen Sie Ihre Entscheidung!*

2. *Überlegen Sie, ob Randklassen nötig sind.*

3. *Sehen Sie sich das Diagramm in http://populationpyramid.net/ an. Welche Variablen mit welchem Wertebereich liegen hier vor? Welche Messwertklassen gibt es? Gibt es Randklassen? Warum / Warum nicht?*

8 Häufigkeit und Häufigkeitsverteilung

Die *Häufigkeit H* oder $f(x_i)$ innerhalb einer Stichprobe n bezeichnet die Anzahl der gleichen Ausprägungen einer Variablen innerhalb der Urliste. In der Übungstabelle gibt es genau 6 Personen, die 18 Jahre alt sind. Die Ausprägung „18“ der Variable „Alter“ kommt also 6 Mal vor. Die Zahl 6 bezeichnet die *absolute Häufigkeit*. Bezieht man die Häufigkeit einer Ausprägung auf die Gesamtheit der untersuchten Einheiten, erhält man die *relative Häufigkeit*, die als Zahl oder in Prozent angegeben werden kann.

Die Häufigkeiten aller Ausprägungen der Variablen ergeben die Häufigkeitsverteilung einer Variablen. Von Hand kann man die Häufigkeitsverteilung über eine Strichliste ermitteln. Wenn Werte zu Messwertklassen zusammengefasst sind, werden die Häufigkeiten der einzelnen Werte zu einer Klassenhäufigkeit addiert.

Beispiel: In der Übungstabelle gibt es 6 Achtzehnjährige, 1 Siebzehnjährigen und 10 Neunzehnjährige. Die Klassenhäufigkeit der Klasse [17-19] beträgt demnach 17.

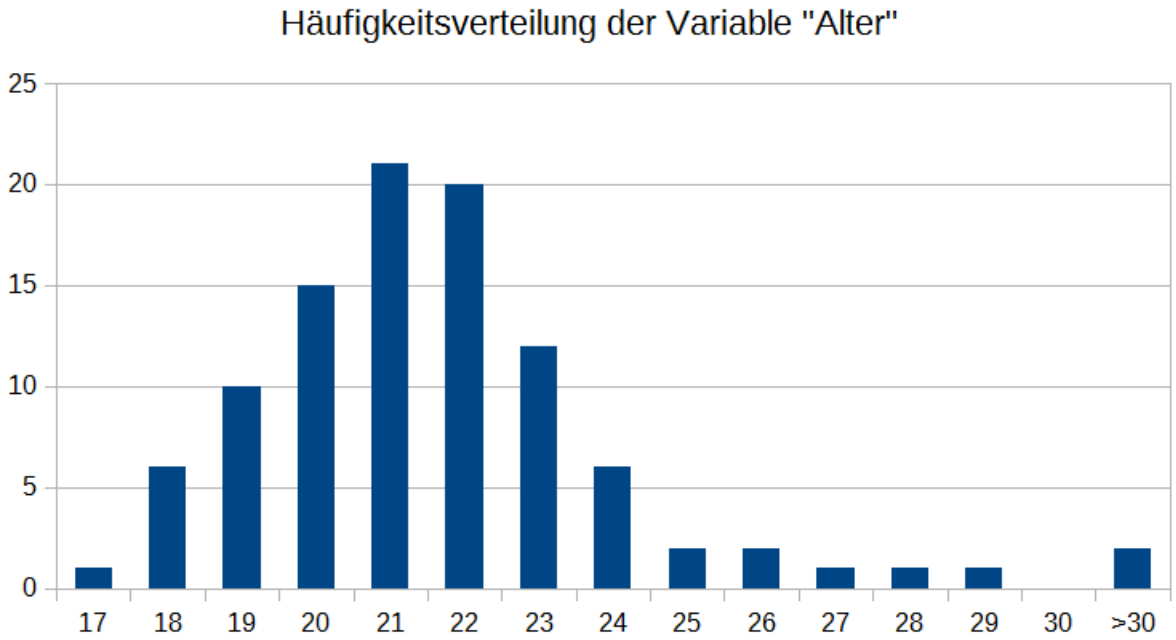
Aufgaben:

Ermitteln Sie anhand der Übungstabelle im Anhang die absolute und relative Häufigkeit

1. *der männlichen und der weiblichen Studierenden,*
2. *der 26-jährigen,*
3. *der Studierenden, die zwischen 500 und 1000 Euro ausgeben (Klassenhäufigkeit).*

Häufigkeitsverteilungen lassen sich sehr gut in Form von Diagrammen veranschaulichen. Dazu benutzen wir ein Koordinatensystem. Wir tragen die Merkmalsausprägungen auf der x-Achse und die Häufigkeiten auf der y-Achse ab. Wenn wir nun die Punkte durch senkrechte Linien mit der x-Achse verbinden, erhalten wir ein Stabdiagramm. Wenn man statt der dünnen Linien ausgefüllte Säulen zeichnet, nennt man das ein Säulendiagramm, wobei die Breite der Säule unerheblich ist. Dieser Begriff wird in der Literatur allerdings auch für Histogramme benutzt, wo die Säulenbreite sehr wohl eine Bedeutung hat. (in unserer Beispieltabelle jedoch nicht)

Da es sich bei *Alter* um eine Verhältnisskala handelt, sollte die x-Achse aufsteigend skaliert sein.



Noch nicht so recht verstanden? Ein sehr einfaches Beispiel können Sie sich hier ansehen: http://psychologie.fernuni-hagen.de/methoden/ils/kurs/kap01/kap01_0.html

8.1 Absolute, relative und kumulative Häufigkeit

Häufigkeiten können in verschiedener Form dargestellt werden.

Die absolute Häufigkeit *H* oder h_{abs} eines Wertes oder einer Klasse ist eine Zahl. Sie beschreibt, wie oft ein Merkmal oder eine Merkmalsklasse vorkommt.

Die relative Häufigkeit *h* oder h_{rel} bezeichnet den Anteil eines Wertes oder einer Klasse an der Gesamtzahl der Werte. Sie kann als Bruch oder in Prozent ausgedrückt werden und wird folgendermaßen berechnet:

$$h = \frac{H}{n}(alsBruch); h = \frac{H}{n} \cdot 100(Prozent)$$

Wenn man jede Häufigkeit zu der jeweils vorigen addiert, erhält man die *kumulierte Häufigkeitsverteilung*.

Als Beispiel betrachten wir das Lebensalter von 20 Studienanfängern. Die geordnete Werteliste ergibt:

17; 17; 18; 18; 18; 19; 19; 19; 20; 20; 20; 20; 20; 21; 21; 21; 22; 22; 23; 24

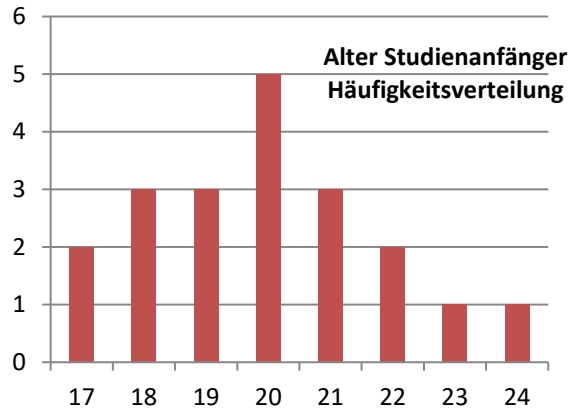
Die Häufigkeiten lassen sich hier relativ leicht ablesen. Daraus können wir eine Tabelle machen.

Alter	H	h [%]	H(kumuliert)	h(kumuliert)
17	2	10	2	10
18	3	15	5	10+15= 25
19	3	15	8	25+15= 40
20	5	25	13	40+25= 65
21	3	15	16	65+15= 80
22	2	10	18	80+10= 90
23	1	5	19	90+5= 95
24	1	5	20	95+5= **100**
Summe:	**20**	**100**		

Kontrollsummen:
Die Summe aller absoluten Häufigkeiten muss *n* sein.
Die Summe aller relativen Häufigkeiten ist 1 (Bruch) bzw. 100 (Prozent).

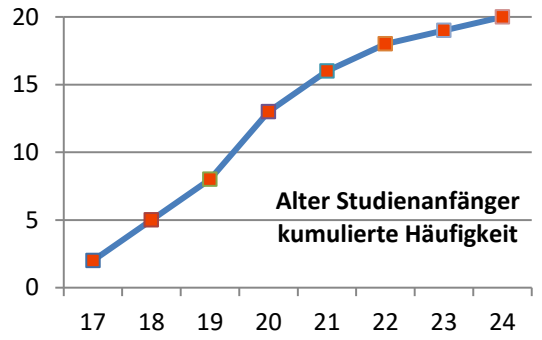
Aufgabe:

Woran erkennen Sie die Richtigkeit der kumulativen Häufigkeitsberechnung?



Der Nutzen dieser Berechnungen wird klar, wenn man aus diesen Zahlen Diagramme erstellt. Für die absolute und die relative Häufigkeit sehen die Diagramme exakt gleich aus, lediglich die Beschriftung der senkrechten Achse ist unterschiedlich. Man kann an der Höhe der Balken sofort sehen, welche Werte besonders oft vorkommen.

***Frage**: Wie würde das Diagramm aussehen, wenn von jedem Jahrgang gleich viele Studenten existieren?*



Die grafische Darstellung der kumulativen Häufigkeitsverteilung ist gewöhnungsbedürftig. Da die Werte immer hinzuaddiert werden, steigt die Kurve immer an. Bei kleinen Häufigkeiten ist sie flach, bei großen Häufigkeiten steigt sie steiler und es gibt einen „Buckel“.

***Frage**: Wie würde das Diagramm rechts aussehen, wenn es von jedem Jahrgang genau drei Studenten gäbe?*

9 Lagemaße (Maße der zentralen Tendenz)

Die Maße der zentralen Tendenz sind Skalenwerte, die Aufschluss darüber geben, welcher Skalenwert die verschiedenen Messwerte am besten „repräsentiert“.

9.1 Minimum / Maximum (x_{min} / x_{max})

Das Minimum ist der kleinste vorkommende Wert einer Variablen. Das Minimum kann man durch Sortieren herausfinden.

9.2 Modus oder Modalwert (x_{mod})

Der *Modus* (*Modalwert*) einer Verteilung ist derjenige Skalenwert, der am häufigsten auftritt. Wenn mehrere Skalenwerte zu Klassen zusammengefasst werden, gilt die Klassenmitte der häufigsten Klasse als Modalwert der Verteilung. Der Modus ist das einzig sinnvolle Maß der zentralen Tendenz bei Nominalskalen. Es kann jedoch vorkommen, dass mehrere Werte in einer Skala gleich häufig vorkommen. In diesem Fall gibt es mehrere Modi. Man spricht dann von einer multimodalen oder mehrgipfligen Verteilung.

9.3 Mittelwert ($\bar{x}$)

Der Mittelwert oder das arithmetische Mittel $\bar{x}$ ist die Summe der Skalenwerte geteilt durch die Anzahl der Messwerte.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Aufgaben:

1. *Überlegen Sie, für welche Variablen der Übungstabelle welche Maße sinnvoll ermittelt werden können und ermitteln Sie dann:*
   das Minimum:
   das Maximum:
   dem Modalwert:
   den Mittelwert

9.4 Median ($x_{50} oder x_M$)

Der *Median* (*Zentralwert*) einer Verteilung ist derjenige Wert einer Skala, der die Menge der Messwerte in genau zwei gleich große Hälften teilt, in eine *untere Teilliste* und eine *obere Teilliste*. Bei einer ungeraden Zahl ist der Messwert genau der Wert, der in der Mitte liegt. Bei einer geraden Zahl von Messwerten ist der Median das arithmetische Mittel zwischen den beiden mittleren Werten.

Beispiele:

n = 13 (ungerade)

2	4	6	8	10	12	14	16	18	20	22	24	26
						x_{50}						

n = 12 (gerade)

3	6	9	12	15	18	21	24	27	30	33	36
					$x_{50} = \frac{x_6 + x_7}{2} = \frac{18 + 21}{2}$						

Bei einer ungeraden Zahl der Werte entspricht der Median also genau dem einen Wert in der Mitte. Bei geradem *n* ist der Median das arithmetische Mittel der benachbarten Werte. Es gilt:

$$x_M = \begin{cases} x_{\frac{n+1}{2}} & \text{für } n \text{ ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & \text{für } n \text{ gerade} \end{cases}$$

9.5 Quantile (x_p)

Quantile sind eine spezielle Form der Lagemaße.

Allgemeine Definition: *Für einen Prozentsatz p ist das p-Quantil einer Verteilung definiert als die Zahl, unterhalb derer p der Werte und oberhalb der 1 - p der Werte liegen.*

Der Median ist das 50%-Quantil. Das 75%-Quantil und das 25%-Quantil einer Verteilung werden als *Quartile* bezeichnet. Das *erste Quartil* (x_{25}) teilt die untere Teilliste in zwei Hälften. Es ist also sozusagen der Median der unteren Teilliste. Das *dritte Quartil* (x_{75}) teilt die obere Hälfte in der Mitte.

Mit dem Median kann man nur erkennen, ob sich die überwiegende Zahl der Werte oder auch Ausreißer[2] in der linken oder rechten Hälfte der Wertmenge befinden. Um die Lage der Ausreißer genauer angeben zu können, teilt man die Menge der Skalenwerte nicht wie beim Median in zwei, sondern in vier Teile und erhält damit 5 Werte. Dabei ist der erste Wert (Quartil 0) das Minimum, das 2. Quartil identisch mit dem Median und das 4. Quartil entspricht dem Maximum. Berechnet werden müssen eigentlich nur das 1. und 3. Quartil.

Quartil	Bezeichnung	Eigenschaft
0.	Minimum - x_{min}	
1.	1. Quartil - x_{25}	25% der Daten
2.	Median - x_{50}	50% der Daten
3.	3. Quartil - x_{75}	75% der Daten
4.	Maximum - x_{max}	100% der Daten

Wenn nur wenige Daten vorliegen, muss man bei der Berechnung der Quartile beachten, dass vor und nach den Quartilen nicht die gleiche Anzahl von Daten ist. Unterhalb von x_{25} befinden sich z.B. nur 25% der Werte, darüber jedoch 75%. (s. Tabelle unten) Bei sehr vielen Werten spielt dies jedoch keine Rolle. Hier kann man die Quartile einfach als Mittelwert der benachbarten Zahlen berechnen.

2 Als Ausreißer bezeichnet man meist Werte, die weiter als das 1,5-fache des Quartilsabstands von Q1 bzw. Q3 entfernt sind. Es gibt allerdings auch andere Definitionen.

Statistikfunktionen in Tabellenkalkulationsprogrammen (z.B. Excel, Calc) nutzen verschiedene Methoden, so dass auch minimal abweichende Werte entstehen können.

Hier die penible Methode:

Wert Nr	1	2	3	4	5	6	7	8	9	10	11	12
Wert	3	7	12	15	20	22	39	40	58	62	70	204
Quartile	Q_0=MIN | **Q_{25}=12·0,25+15·0,75=14,25** | Q_{50}=(22+39)/2 =Median | **Q_{75}=58·0,75+62·0,25 = 59** | Q_{100}=MAX											
einfache Berechn.	Q_1 =(12+15)/2 =13,5 | | Q_3 =(58+62)/2 =60 |											

Die Differenz zwischen x_{75} und x_{25} heißt Quartilsabstand (engl: Interquartile Range – IQR). Zwischen x_{25} und x_{75} liegen die mittleren 50% der Werte. Quartile und Quartilsabstand sind robuste Maße, da sie durch Ausreißer nicht verändert werden.

Quintile sind Merkmalsklassen, die entstehen, wenn man die Menge der Skalenwerte genau in fünf Teile zu 20% teilt. Es gelten dieselben Regeln wie bei der Klassenbildung. Quintile werden z.B. zur Messung der Einkommensungleichheit in einer Bevölkerung benutzt. Dazu berechnet man den Abstand zwischen den Einkommen des oberen Fünftels und des unteren Fünftels der Bevölkerung (Quintilsabstand).

Man kann Quantile allgemein mit folgender Methode berechnen:

$$Q_p \approx x_{[n\cdot p]+1} \ \ f\ddot{u}r(n\cdot p)\ nicht\ ganzzahlig$$

[ ]=Gauss-Klammer=ganzzahliger Anteil der Zahl, z.B. [8,265]=8

$$Q_p \approx \frac{1}{2}\left(x_{n\cdot p} + x_{n\cdot p+1}\right) f\ddot{u}r(n\cdot p) ganzzahlig$$

***Aufgaben**: Gegeben sei folgende geordnete Werteliste*:

4	8	13	16	20	22	28	40	60	62	70	79

Ermitteln Sie folgende Lagemaße: Minimum, Maximum, Modalwert (falls vorhanden), Median, Quartile und den Quartilsabstand. Ermitteln Sie die Quantile durch Abzählen sowie durch die Anwendung der allgemeinen Formeln.

9.6 Der Nutzen von Mittelwert, Median und Quartilen

In den Mittelwert fließen alle Messwerte ein, in den Median aber nur die beiden Werte in der Mitte der Liste. Der Median wird dazu benutzt, um „Ausreißer“ zu neutralisieren. Ab wann ein Wert als Ausreißer betrachtet wird, wird unterschiedlich definiert. Am häufigsten definiert man sie als Werte außerhalb der doppelten oder der dreifachen Standardabweichung. Sie deuten häufig auf grobe Fehler der Datenerfassung hin. Zur Verdeutlichung wählen wir aus zwei Gruppen von Probanden willkürlich aus:

Proband Nr.	Ausgaben	Proband Nr.	Ausgaben
1	370	1	370
2	420	2	420
3	450	3	450
4	500	4	500
5	542	5	542
6	580	6	580
7	600	7	600
8	700	8	700
9	707	9	2300
Mittelwert	541	Mittelwert	718
Median	542	Median	542
Differenz	1	Differenz	177

Wie man sieht, wird der Mittelwert durch den einen Probanden, der 2300 Euro zur Verfügung hat, stark nach oben verändert. Wenn man also (was in vielen Statistiken in der Zeitung gemacht wird), den Mittelwert als Maß nimmt, könnte man meinen, dass nun alle Probanden mehr Geld hätten.

Der Vergleich der Zentralwerte zeigt aber, dass sich die Situation der meisten Probanden nicht verändert hat. Eine hohe Differenz zwischen Mittelwert und Median deutet darauf hin, dass es Ausreißer gibt.

Als Ausreißer werden Werte bezeichnet, die sehr stark von der Mehrheit der Werte abweichen. Ab wann ein Wert als Ausreißer bezeichnet wird, ist in der Literatur unterschiedlich. Die häufigste Definition ist das dritte Quartil plus der 1,5-fache Quartilsabstand.

Aufgaben:

1. *In einem kleinen Dorf gibt es folgende Einkommensverteilung (in Euro/Monat):*

Familie 1	Familie 2	Familie 3	Familie 4	Familie 5
1400	1550	1200	1150	1600

1.1. Berechnen Sie Median und Mittelwert.

Nun baut eine reiche Familie ein neues Haus im Dorf. Die Verteilung ändert sich wie folgt:

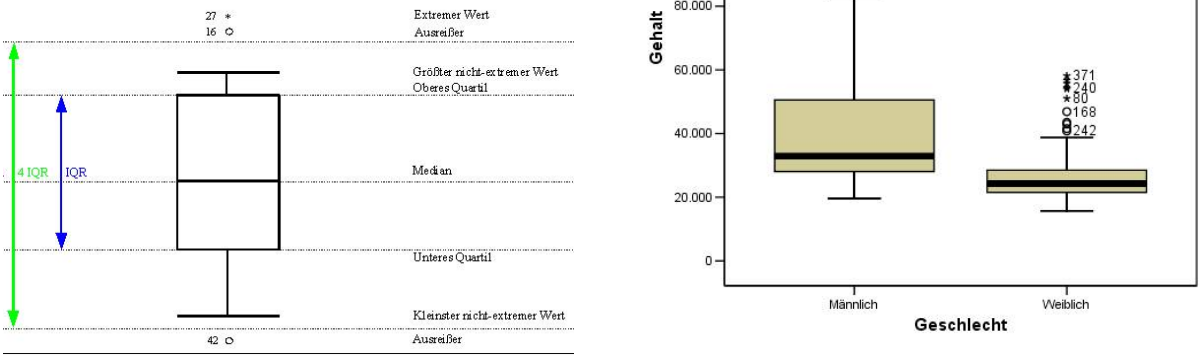
Familie 1	Familie 2	Familie 3	Familie 4	Familie 5	Familie 6
1400	1550	1200	1150	1600	26000

1.2. Ermitteln Sie Mittelwert, Median und Quartile.
1.3. Was kann man an den Zahlen erkennen?
1.4. Kann man aufgrund der Zahlen sagen, dass es den Menschen in dem Dorf nun besser geht?

9.7 Box-Plot-Diagramme

Um die Verteilung von Werten auf einen Blick sichtbar zu machen, benutzt man oft so genannte Boxplot-Diagramme, auch Box and Whiskers genannt. Die einfachste Methode zum Erstellen eines Boxplot benutzt Minimum, Q1, Median, Q3 und Maximum. Dabei werden die Werte zwischen dem ersten und dritten Quartil als Rechteck (Box) gezeichnet. Der Median befindet sich als Strich innerhalb dieser Box. Dünne Striche (Whiskers) gehen nach unten und nach oben bis zu einem definierten Wert bzw. bis zu Minimum und Maximum.

Viele Darstellungen zeichnen den Whisker nicht bis zum Minimum bzw. Maximum sondern bis zu einem willkürlich definierten Wert (z.B. bis zum 1,5-fachen oder vierfachen Quartilsabstand, s. Bild). Werte außerhalb dieses Bereichs werden in diesem Fall mit Punkten als Ausreißer dargestellt.



Boxplot allgemein (links) und der berühmte „Gender Pay Gap“ (rechts), mit dem viel statistischer Unsinn getrieben wird
Quelle: Marktforschungs-Wiki, http://marktforschung.wikia.com/wiki/Box-Plot

Beispiel:
Um den Benzinverbrauch seines Autos zu kontrollieren, schreibt Herr A auf, wie viel Liter sein Auto pro 100 km verbraucht:

6,5 ; 7,4 ; 7,4 ; 7,8 ; 6,7 ; 7,3 ; 6,7 ; 7,3 ; 6,7 ; 7,6 ; 6,4 ; 7,5 ; 6,5 ; 6,9 ; 7,8 ; 7,2 ; 6,9 ; 6,7 ; 7,6 ; 7,4

Lösung:

1. Ordnen der Daten
6,4; 6,5; 6,5; 6,7;6,7; 6,7; 6,7; 6,9; 6,9; 7,2; 7,3; 7,3; 7,4; 7,4; 7,4; 7,5; 7,6; 7,6; 7,8; 7,8

2. Bestimmen des Medians
6,4; 6,5; 6,5; 6,7;6,7; 6,7; 6,7; 6,9; 6,9; **7,2;|7,3**; 7,3; 7,4; 7,4; 7,4; 7,5; 7,6; 7,6; 7,8; 7,8

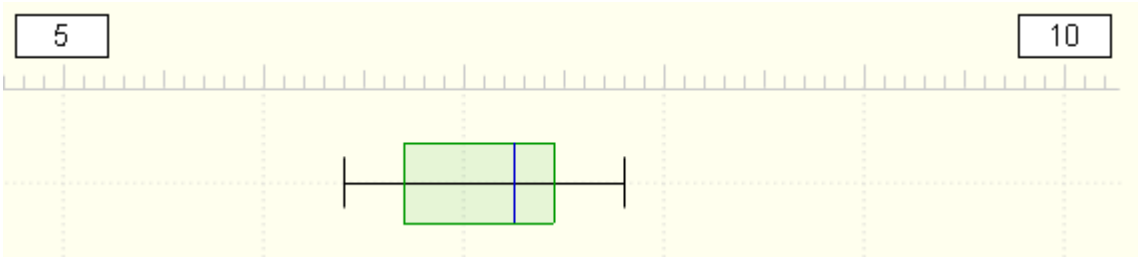
$$\frac{7{,}2 + 7{,}3}{2} = 7{,}25$$

3. Bestimmen der Quartile (hier auf einfache Weise):

6,4; 6,5; 6,5; 6,7;**6,7;|6,7**; 6,7; 6,9; 6,9; 7,2; | 7,3; 7,3; 7,4; 7,4; **7,4;|7,5**; 7,6; 7,6; 7,8; 7,8

$$Q_{25} = \frac{6{,}7 + 6{,}7}{2}; Q_{75} = \frac{7{,}4 + 7{,}5}{2} = 7{,}45$$

4. Boxplot zeichnen



Quelle: http://www.mathe-trainer.de/Klasse8/Wahrscheinlichkeitsrechnung/Block1/Loesungen/A1-1.htm

Box-Plot-Diagramme kann man übrigens mit den gängigen Office-Programmen nur sehr schwer oder gar nicht erzeugen. Es gibt aber eine große Auswahl an Online-Tools, die dies können, z.B. (EMBACHER)

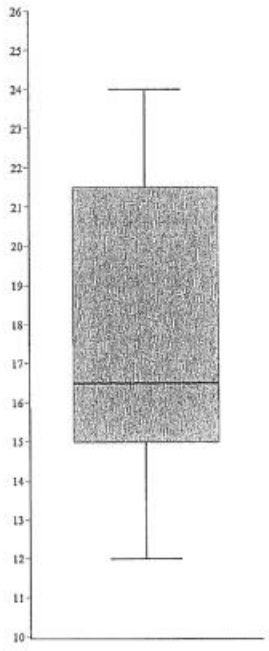


Abbildung 1: Wegzeiten in Minuten

Aufgabe:

Eine Studentin hat sich drei Wochen lang die Zeit notiert, die sie für den Weg zur Uni braucht. Der Boxplot zeigt die Ergebnisse.

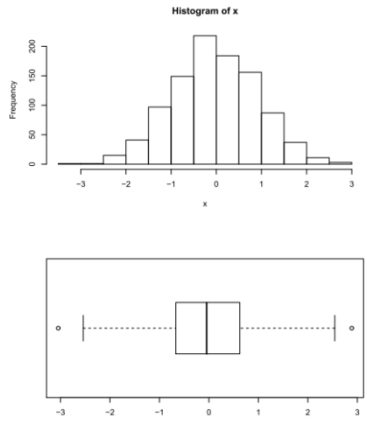
1. Wo liegen bei diesen Daten das Minimum und das Maximum? Wie groß ist der Median und der Quartilsabstand? Geben Sie auch das erste und das dritte Quartil der Daten an. (Ungefähre Werte ablesen!)

2. (Für Profis) Welche Vermutungen können Sie aus diesen Maßen ableiten? Welche Fragen würden Sie der Studentin stellen um diese Vermutungen zu klären?

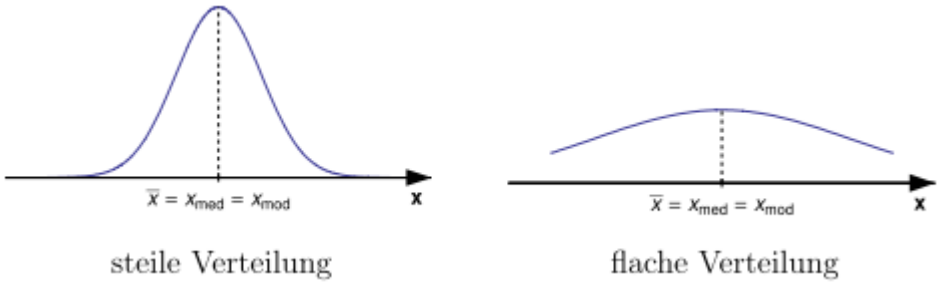
Dieses Beispiel stammt übrigens aus einer Original-Statistik-Klausur für Pädagogik-Studenten der TU Darmstadt.

9.8 typische Verteilungen

Die Art der Verteilung von Daten kann man an dem Histogramm der Häufigkeiten oder an dem entsprechenden Boxplot ablesen. Den Typ einer Verteilung kann man außerdem durch den Vergleich von Median, Modalwert und Mittelwert ermitteln (Fechnersche Lageregel).Typische Verteilungen kann man nach mehreren Kriterien einteilen.



Eine annähernd symmetrische Verteilung. In der Sozialwissenschaft sind solche Verteilungen immer nur annähernd symmetrisch.



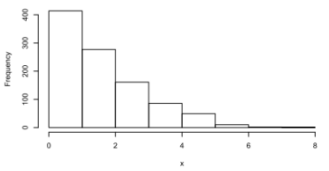
steile Verteilung flache Verteilung

Bei einer symmetrischen Verteilung sind die Häufigkeiten rechts und links vom Mittelwert gleich verteilt. Bei einer steilen Verteilung sind die Unterschiede zwischen den Werten sehr groß. bei einer flachen Verteilung sind sie klein. Es gilt:

$$\acute{x} = x_{50} = x_{mod}$$

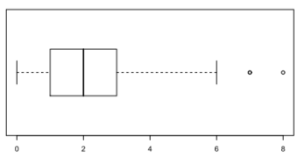
($\acute{x}$=Mittelwert; x_{50}=Median; x_{mod}=Modalwert)

Typische symmetrische Verteilungen sind z.B. die Körpergröße bei Männern oder Frauen.



Links sehen Sie eine nicht symmetrische Verteilung. Die kleinen Werte haben eine hohe Häufigkeit, die großen Werte haben eine niedrige Häufigkeit. Man nennt diese Verteilung *rechtsschief*.

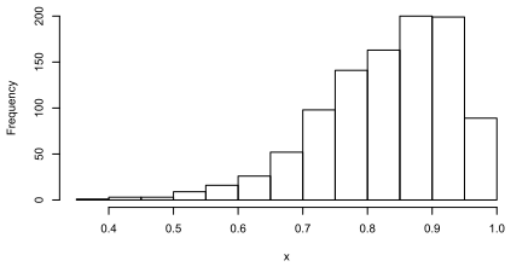
Es gilt: $x_{mod} < x_{50} < \acute{x}$



Ein typisches Beispiel für rechtsschiefe Verteilungen sind Einkommensverteilungen. Während der größte Teil der Bevölkerung kleine oder mittlere Einkommen bezieht, gibt es nur wenige, die hohe oder sehr hohe Einkommen haben.

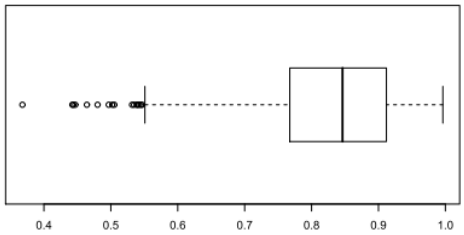
In der folgenden Abbildung ist die Verteilung der Werte *linksschief*. Das bedeutet, dass die Mehrheit der Werte rechts vom Median liegt und dass die Häufigkeit nach links abnimmt.

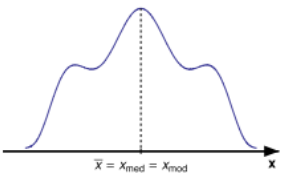
Es gilt:



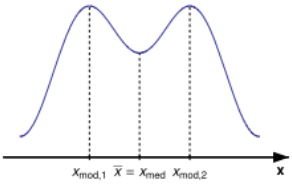
$$\acute{x} < x_{50} < x_{mod}$$

Eine solche Verteilung ergibt sich z.B., wenn man die Leistungen von Profisportlern misst. Die meisten Profisportler sind in der Nähe der menschlichen Leistungsgrenze. Profisportler, die durchschnittliche oder sogar schlechte Leistung haben, sind dagegen eher selten.





unimodale Verteilung



bimodale Verteilung

Bei einer unimodalen (oder monomodalen) Verteilung gibt es nur einen Modalwert, bei einer bimodalen Verteilung gibt es zwei und bei einer multimodalen Verteilung gibt es mehr als zwei Modalwerte. Wenn wir z.B. die Körpergröße von Männern und Frauen getrennt messen, wird sich wahrscheinlich jeweils eine unimodale Verteilung ergeben. Messen wir beide zusammen, ergibt sich wahrscheinlich eine bimodale Verteilung mit einem Modalwert für die häufigste Körpergröße von Frauen und einem Modalwert für die häufigste Körpergröße von Männern.

10 Dispersionsmaße (Streuungsmaße)

Die Streuungsmaße geben Hinweise darauf, wie die Werte verteilt sind und wie stark sie um den Mittelwert streuen.

10.1 Spannweite / Range

$$R = x_{max} - x_{min}$$

Die Spannweite ist das absolute Maß der Streuung einer Stichprobe.

Aufgabe: Nennen Sie die Spannweite der Merkmale in der Übungstabelle. Für welche Merkmale kann man keine Spannweite ermitteln?

10.2 Quartilsabstand

Der Quartilsabstand *Q* ist die Differenz zwischen dem Skalenwert nach 25% (x_{25}) und nach 75% (x_{75}) der Merkmalsträger. Er gibt also den Wertebereich an, in dem die mittleren 50% der Werte liegen. Der Quartilsabstand gilt als robustes Maß. Da er nicht vom Mittelwert abhängt, wird er von extremen Ausreißern nicht oder kaum beeinflusst.

Aufgabe: Berechnen Sie den Quartilsabstand der Variable „Alter“ in der Übungstabelle.

10.3 Durchschnittliche Abweichung (AD) / mittlere Abweichung

Das ist die Summe der Beträge der Abstände aller Messwerte zum arithmetischen Mittel $\bar{x}$ geteilt durch die Anzahl der Messwerte n. Würden wir statt der Beträge die Abstände selbst nehmen, ergäbe die Summe 0. Die mittlere Abweichung ist ähnlich der Spannweite ein Maß für die Streuung der Messwerte.

	Werte	Abstand a	Beträge	a^2
Wert 1	4	-1	1	1
Wert 2	5	0	0	0
Wert 3	7	2	2	4
Wert 4	3	-2	2	4
Wert 5	6	1	1	1
Summe	25	0	6	10
	MW = 5		AD = 1,2	$s^2 = 2$

Beispiel oben: Die Summe der Variablenwerte 4,5,7,3,6 ergibt 25. Die Summe dividiert durch die Anzahl der Werte *n* ergibt den Mittelwert 5. Die Abstände vom Mittelwert stehen in der zweiten Spalte. Ihre Summe ist 0. Die Beträge der Abstände haben keine Vorzeichen. Ihre Summe ist 6.

10.4 Varianz (V[x]; s^2; σ^2)

Die Varianz ergibt sich aus der Abweichung der einzelnen Messwerte vom Mittelwert. Wenn man die Abweichungen vom Mittelwert quadriert, bekommen sie ein größeres Gewicht. Außerdem werden beim Quadrieren automatisch alle Abstände positiv.

Definition: *Die Varianz ist das arithmetische Mittel der Summe aller quadrierten Abstände zwischen Messwerten und dem arithmetischen Mittel.*

Dabei unterscheidet man zwei Fälle. Bei der Berechnung der Varianz aus der gesamten Population teilt man die Summe der quadrierten Abweichungen durch n, im Fall einer Stichprobe teilt man durch n-1.

Varianz s^2, wenn alle Daten bekannt sind (empirische Varianz):

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Varianz σ^2 einer Stichprobe:

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Sind alle Messwerte gleich, so wird die Varianz 0. Die Varianz oder Streuung gibt also an, ob die Messwerte eng um den Mittelwert liegen (kleine Varianz) oder ob es viele, eventuell extreme Abweichungen vom Mittelwert gibt (große Varianz/Streuung).

In der folgenden Tabelle wird die quadratische Abweichung gezeigt:

i	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	4	-1	1
2	5	0	0
3	7	2	4
4	3	-2	4
5	6	1	1
n=5 *Summe*	*25*	*0*	*10*

Die dritte Spalte zeigt die Quadrate der Abweichungen. Ihre Summe ist 10. Wenn alle Daten bekannt sind, dividiert man diese Summe durch *n*=5. Es ergibt sich eine Varianz von 2. Bei einer Stichprobe dividiert man durch n-1, also durch 4. In diesem Fall ergibt das eine Varianz von 2,5. Bei sehr großen *n* ist der Unterschied zwischen beiden Berechnungsarten unerheblich.

Standardabweichung (s oder σ)

Die Standardabweichung (Streuung im engeren Sinne) ist die Wurzel aus der Varianz. Sie ist „handlicher“, weil sie auf derselben Einheit wie die Messwerte und nicht auf deren Quadraten beruht.

Auch bei der Berechnung der Standardabweichung muss man unterscheiden, ob alle Daten vorliegen oder ob es sich um eine Stichprobe handelt.

Standardabweichung der Grundgesamtheit (empirische Standardabweichung):

$$s = \sqrt{s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Standardabweichung der Stichprobe:

$$\sigma = \sqrt{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

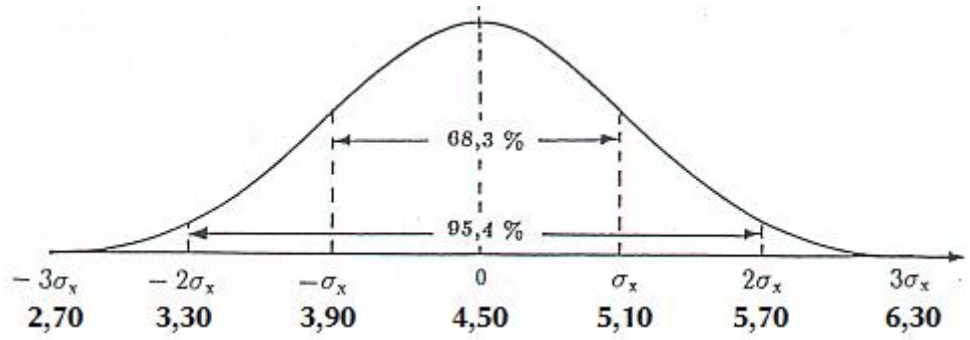
Halten wir fest: Alle drei hier genannten Streuungsmaße zeigen dasselbe, nämlich die durchschnittliche Entfernung aller gemessenen Werte vom Mittelwert. In der Praxis wird fast immer die Standardabweichung angegeben. Warum dies so ist, sehen Sie im folgenden Abschnitt.

10.5 Was sagt uns die Standardabweichung?

Im Zusammenhang mit den Quartilen und dem Quartilsabstand haben wir bereits diskutiert, was diese Größen über die Verteilung der Werte aussagen. Der Abstand zwischen Q_{25} und Q_{75} sagt aus, wie weit die Werte um den Median (Q_{50}) gestreut sind. Der zentrale Wert ist hierbei der Median. Der zentrale Wert der Standardabweichung ist hingegen der Mittelwert. Zum besseren Verständnis betrachten wir eine ideale Häufigkeitsverteilung, die so genannte *Normalverteilung*. Eine solche Verteilung entsteht, wenn Messwerte zufällig um einen Mittelwert streuen.

Nehmen Sie an, Sie befragen in der Mensa 1000 Studenten, wie viel Geld sie bezahlt haben. Wir erhalten einen Mittelwert von 4,50 Euro und eine Standardabweichung σ von 0,6. Wenn die

Auswahl der Befragten zufällig ist, könnte sich dann folgende Häufigkeitsverteilung ergeben (fiktives Beispiel):



Der Mathematiker Carl Friedrich Gauß (Sie sehen ihn auf dem Titelblatt dieses Skripts) hat diese Verteilung mathematisch beschrieben. Als G-Kurs sparen wir uns die höhere Mathematik und betrachten die Kurve einfach ein wenig genauer. Dabei stellen wir einige interessante Merkmale fest:

- Die Kurve sieht aus wie eine Glocke. Sie heißt deshalb auch *Gauss'sche Glockenkurve*.
- Die Kurve ist symmetrisch zum Mittelwert. Unterhalb und oberhalb des Mittelwerts liegen gleich viele Werte.
- Im oberen Teil ist die Kurve rechts gekrümmt (wie ein Dach) und im unteren Teil sind beide Kurventeile links gekrümmt (wie eine Schüssel). Der Punkt, an dem die Rechtskrümmung jeweils in die Linkskrümmung übergeht, heißt *Wendepunkt*.
- Der x-Wert der beiden Wendepunkte entspricht genau der Standardabweichung.

Es sieht so aus, als würde die Kurve an beiden Enden die x-Achse berühren. Das ist mathematisch nicht der Fall – sie geht bis ins Unendliche weiter. Man sieht aber, dass die Werte dann so klein werden, dass sie für die praktische Statistik uninteressant sind. Deshalb schneiden wir für statistische Zwecke die Kurve rechts und links einfach ab.

Im Fall der Mensa-Befragung sagt uns die Standardabweichung, dass die durchschnittliche Abweichung vom Mittelwert 0,6 Euro, also 60 Cent beträgt. Mit anderen Worten: Die Studenten, die in der Mensa sparsamer sind, geben im Durchschnitt 3,90 Euro aus. Man sieht an der Kurve auch auf den ersten Blick, dass die Werte um den Mittelwert die größte Häufigkeit besitzen.

Die Zahl der Menschen, die mehr bzw. weniger als 4,50 Euro ausgeben, wird immer kleiner, je weiter sich die Größe vom Mittelwert entfernt. Mit der Standardabweichung kann man sogar abschätzen, wie groß diese Zahl ist:

- Im Bereich $-\sigma$ bis $+\sigma$, also zwischen 3,90 Euro und 5,10 Euro, befinden sich ungefähr 68,3% aller Werte. Anders gesagt: 68,3 % der Mensabesucher geben zwischen 3,90 Euro und 5,10 Euro aus.
- Zwischen -2σ und $+2\sigma$ befinden sich sogar 95,4 % der Werte.
- Werte unterhalb und oberhalb von 2σ bezeichnet man als Ausreißer.

In Boxplots (vgl. 9.7) wird als oberes und unteres Ende der *Whiskers* oft nicht Minimum und Maximum angegeben, sondern die Werte zwischen -3σ und $+3\sigma$.

Aufgaben:

1. *Wie viele Studenten haben in dem obigen Beispiel weniger als 3,30 Euro ausgegeben?*
2. *Wie viele Studenten haben mehr als 3,30 Euro und weniger als 5,70 Euro ausgegeben?*
3. *Prüfen Sie, ob die Altersverteilung in Kap.* 8.1 *annähernd normalverteilt ist.*

11 Zusammenhang zwischen Daten

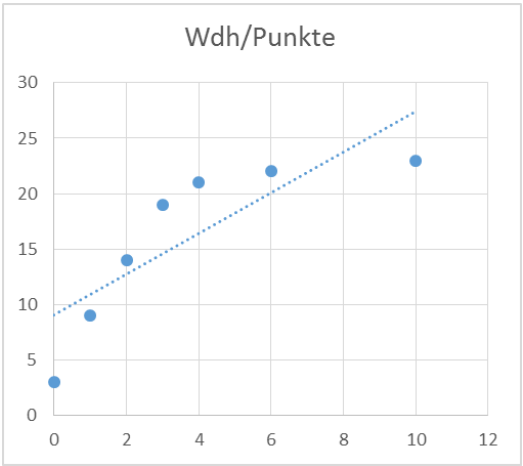
Bisher haben wir einzelne Merkmale statistisch ausgewertet. Oft ist es aber interessant, ob es einen Zusammenhang zwischen mehreren Merkmalen gibt. Am Beispiel unserer Übungstabelle könnte man z.B. fragen, ob ältere Studierende mehr Geld haben als jüngere. In den Worten der Statistik: Wir fragen, ob zwischen der Variable Alter und der Variable Ausgaben eine Korrelation besteht.

11.1 Streuungsdiagramme

Um grob abschätzen zu können, ob zwischen Datenreihen ein Zusammenhang besteht, hilft ein Streuungsdiagramm. In diesem Diagramm entsprechen die beiden Achsen den beiden Variablen, die wir in Beziehung setzen.

Um den Zusammenhang zweier Variablen grafisch darzustellen, konstruiert man ein *Streuungsdiagramm.* Im folgenden Beispiel messen wir, wie oft ein Student den Stoff wiederholt hat (Variable 1) und wie viele Punkte er in der Klausur bekommen hat (Variable 2). Wenn wir beide Variablen auf den Achsen eines Koordinatensystems abtragen, erhalten wir für jede Kombination einen Punkt. Ein statistischer (nicht kausaler!!) Zusammenhang ist wahrscheinlich, wenn die Punkte genau oder in der Nähe einer gedachten Gerade liegen, deren Abstand zu allen Punkten minimal ist. Eine solche Gerade heißt *Ausgleichsgerade, Regressionsgerade* oder *Trendlinie.* Auf die Berechnung der Geradengleichung verzichten wir hier, das überlassen wir dem Computer.

h	P
0	3
1	9
2	14
3	19
4	21
6	22
10	23



Streuungsdiagramm mit Trendlinie

11.2 der Korrelationskoeffizient

Wir können nun bereits sehen, ob ein Zusammenhang zwischen den Datenreihen besteht. Um festzustellen, wie stark dieser Zusammenhang ist, berechnet man den Korrelationskoeffizienten *r*, der immer zwischen -1 und +1 liegt. Wenn kein Zusammenhang zwischen den beiden untersuchten Messreihen bzw. Variablen besteht, ist $r = 0$. Wenn der Zusammenhang vollständig ist, ist $r = 1$ oder $r = -1$, also $|r|=1$

Korrelationen können größer oder kleiner als 0 sein, in der Praxis sind sie das immer. Bei einer positiven Korrelation gehen hohe Werte in der einen mit hohen Werten in der anderen Variablen einher. (z.B.: Körpergröße/Schuhgröße: Größere Menschen haben in der Regel auch ein größere Füße.) Bei einem negativen Zusammenhang ergeben hohe Werte in der einen niedrige Werte in der

anderen Variablen. (z.B. Geschwindigkeit/Zeit: Je größer die Geschwindigkeit ist, desto kleiner ist die Zeit, die ich für eine bestimmte Strecke brauche.) Im Rahmen der empirischen Sozialwissenschaften ist der *Pearsonsche Produkt-Moment-Korrelationskoeffizient* am gebräuchlichsten bzw. am wichtigsten. Dieser wird folgendermaßen berechnet:

$$r_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right) \cdot \left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

Ein einfaches Beispiel

x_i seien Werte der Variablen x, y_i seien Werte der Variablen y und z_i seien Werte der Variablen z (z.B. die Punkte in verschiedenen Klausuren)

Zunächst prüfen wir, ob Variable x und Variable y korrelieren.

i	**x_i**	**y_i**	**x_i-$\bar{x}$**	**y_i-$\bar{y}$**	**(x_i-$\bar{x}$)* (y_i-$\bar{y}$)**
1	2	4			
2	4	6			
3	6	8			
4	8	10			
5	10	12			
	$\bar{x}$	$\bar{y}$			
Mittelw.	30/5=6	40/5=8			
σ	2,83	2,83			

***Aufgabe**: Berechnen Sie die fehlenden Werte und tragen Sie diese ein.*
Nun prüfen wir die Korrelation der Variable x mit der Variable z

i	**x_i**	**z_i**	**x_i-$\bar{x}$**	**z_i-$\bar{z}$**	**(x_i-$\bar{x}$)* (z_i-$\bar{z}$)**
1	2	10			
2	4	8			
3	6	6			
4	8	4			
5	10	2			
	$\bar{x}$	$\bar{z}$			
	30/5=6	40/5=8			
Standardabw.	2,83	2,83			

Der Korrelationskoeffizient ist ein Maß für die Bedeutsamkeit (=Signifikanz) eines Zusammenhangs zwischen zwei Variablen. Dabei gilt als Faustregel:

|r| < 0,4 niedrige Signifikanz
|r| > 0,4 < 0,7 mittlere Signifikanz
|r| > 0,7 hohe Signifikanz

Aufgaben:

Die folgende Tabelle zeigt Ergebnisse von Klausuren in einem (fiktiven!) G-Kurs.

St. Nr.	Note_DaF	Note_SOWI	Note_Geschichte	Note_Statistik
1	1	2	1	5

2	1	2	2	5
3	2	3	3	4
4	2	3	4	4
5	3	4	5	3
6	3	4	1	3
7	4	5	2	2
8	4	5	3	2
9	5	6	4	1
10	5	6	5	1

1. *Ermitteln Sie zur Wiederholung noch einmal von allen Variablen die Lagemaße und die Standardabweichung.*
2. *Wie viele Korrelationen sind möglich?*
3. *Ermitteln Sie alle Korrelationskoeffizienten. Wie hoch ist jeweils die Signifikanz?*
4. *Beschreiben Sie die jeweiligen Korrelationen sprachlich. (Tipp: Wenn-dann / Je-desto)*
5. *Wenn Sie einen PC zur Verfügung haben, stellen Sie die Werte als Diagramm dar. Was fällt auf?*

12 Korrelation und Kausalität

Sport ist gesund. Broccoli verhindert Krankheiten. Klavierunterricht verhindert, dass Kinder später kriminell werden. Solche Aussagen liest man ständig in den Zeitungen oder hört sie im Fernsehen. In den meisten Fällen beruft man sich dabei auf eine *empirische Studie*, die eine signifikante Korrelation ergeben hat. In den hier zitierten Fällen könnte man z.B. korrelieren:

- Sportliche Aktivität [h/Woche] / Gesundheitszustand [Likert-Skala]
- Aufnahme von Broccoli [g/Woche] / bei Arbeitgeber als krank gemeldete Tage [Tage]
- Teilnahme an Klavierunterricht [ja/nein] / Verurteilung wegen einer Straftat [ja/nein]

Der Zusammenhang von Musikunterricht und Kriminalität überrascht zwar ein wenig, aber die Behauptung, dass Sport gesund sei, leuchtet doch unmittelbar ein, oder? Die Korrelation suggeriert uns, dass der Sport die *Ursache* der Gesundheit ist. Denken Sie aber bitte noch einmal kurz nach:

Kranke Menschen können oder wollen keinen Sport treiben. Könnte nicht auch das die Ursache der Korrelation sein?

Menschen, die viel Gemüse essen, die alle Ernährungsratgeber lesen, achten sehr auf ihre Gesundheit, rauchen nicht und leben auch sonst sehr gesund. Könnte nicht auch das die Ursache für die Korrelation sein und nicht der Broccoli?

Menschen, die Klavierunterricht hatten, kommen in der Regel aus höheren Gesellschaftsschichten (Arme Leute haben kein Klavier.) und werden alleine schon deshalb weniger kriminell.

Halten wir also fest: In Wirklichkeit sagt eine Korrelation gar nichts, aber auch wirklich gar nichts über einen kausalen Zusammenhang aus!

Mehr zum Thema „Lügen mit Statistik“ finden Sie bei (BOSBACH & KORFF, 2011), zusammengefasst bei (GÖTTMANN, 2012).

Aufgabe:

In der Umgebung von Darmstadt gibt es viele Erdbeerfelder. Reife Erdbeeren sind köstlich. Wir gehen regelmäßig aufs Erdbeerfeld und beobachten zwei Variablen:

1. *Die Zahl der reifen Erdbeeren,*
2. *Die Zahl der Menschen, die mit einem T-Shirt bekleidet sind.*

Dies ist das Ergebnis unserer Erhebung:

	reife	Menschen
Monat	Erdbeeren	m. T-Shirt
Januar	0	0
Februar	0	0
März	2	2
April	100	10
Mai	200	20
Juni	1000	100
Juli	1100	110
August	800	80
September	400	40
Oktober	30	3
November	0	0
Dezember	0	0

1. *Berechnen Sie den Korrelationskoeffiizienten (ungefähre Angabe reicht aus) und bestimmen Sie die Signifikanz.*

2. *Welche Aussagen sind richtig?*
   a) Die Menschen tragen T-Shirts, weil es reife Erdbeeren gibt.
   b) Je mehr Menschen ein T-Shirt tragen, desto mehr reife Erdbeeren gibt es und umgekehrt.
   c) Die Erdbeeren werden reif, weil viele Menschen T-Shirts tragen.
   d) In den Monaten, in denen viele Erdbeeren reif werden, tragen auch viele Menschen T-Shirts.
   e) In Monaten, in denen wenige oder keine Menschen T-Shirts tragen, gibt es auch wenige oder keine Erdbeeren.

13 Literatur und Links

BORTZ, JÜRGEN: *Statistik für Human- und Sozialwissenschaftler*. 6. vollst. überarb. u. aktualisierte. Aufl. Berlin : Springer DE, 2005 — ISBN 978-3-540-21271-3

BOSBACH, GERD ; KORFF, JENS JÜRGEN: *Lügen mit Zahlen - Wie wir mit Statistiken manipuliert werden*. München : Heyne Verlag, 2011 — ISBN 978-3-641-05324-6

DIEFENBACHER, HANS ; FRANK, ANDREAS: *Einfach Lernen! Statistik* : Bookboon.com, 2006

DRAKOS, NIKOS ; MOORE, ROSS: *Statistik und Wahrscheinlichkeitsrechnung für InformatikerInnen*. URL http://www.statistik.tuwien.ac.at/public/dutt/vorles/inf_bak/node1.html. - abgerufen am 2013-01-12

DUTTER, RUDOLF: *Statistik*. URL http://www.statistik.tuwien.ac.at/public/dutt/vorles/inf_bak/node1.html. - abgerufen am 2013-01-13

EMBACHER, FRANZ: *Box-Plots*. URL http://www.mathe-online.at/materialien/Franz.Embacher/files/BoxPlot/. - abgerufen am 2015-05-03. — Boxplot

KRÄMER, WALTER: *So lügt man mit Statistik*. 12. Aufl. München : Piper, 2000 — ISBN 978-3-492-23038-4

RAAB-STEINER, ELISABETH ; BENESCH, MICHAEL: *Der Fragebogen - Von der Forschungsidee zur SPSS/PASW-Auswertung*. 2. aktual. Aufl. Parderborn, München : UTB, 2008 — ISBN 978-3-825-28406-0

STATISTA: *Statistik für Anfänger - Grundlagen der Statistik*. URL http://de.statista.com/statistik/lexikon/definition/155/statistik_fuer_anfaenger_grundlagen_der_statistik/. - abgerufen am 2014-09-11. — Statista

14 Anhang

14.1 Anhang 1: Übungstabelle

Datenerhebung unter international Studierenden 2012
(fiktive Werte, Ähnlichkeiten mit echten Daten der TU Darmstadt sind zufällig)

Fragestellung:

1. Wie alt sind Sie? (vollendete Lebensjahre)
2. Sind Sie männlich (m) oder weiblich (w)?
3. Aus welcher Region stammen Sie? (s. Tabelle unten)
4. Wie hoch sind Ihre durchschnittlichen Ausgaben pro Monat?
(für Lebenshaltung und Studium; Angabe in Euro)

Zuordnung der Herkunftsregionen

Region Nr.	Herkunftsregion
1	Westeuropa
2	Osteuropa
3	Nordafika
4	Subsahara-Afrika
5	Asien West
6	Ost- / Südostasien
7	Nordamerika
8	Mittel- / Südamerika

Hinweis zur Arbeit mit der Tabelle:
Wenn Sie mit einer Tabellenkalkulation am Computer arbeiten, benutzen Sie bitte die komplette Tabelle. Sie finden die Tabelle in Moodle. Wenn Sie keinen Computer benutzen, beschränken Sie sich bitte auf die ersten 20 Werte [i_1, …, i_{20}].

i	Alter	Geschlecht	Region	Ausgaben
1	20	w	5	500
2	24	w	8	490
3	22	m	3	350
4	22	m	3	420
5	21	w	8	520
6	24	m	8	300
7	22	w	4	380
8	24	w	8	370
9	23	w	1	660
10	20	w	8	900
11	24	w	8	500
12	22	w	5	521
13	22	m	7	280
14	22	w	5	160
15	22	m	3	420
16	21	m	3	385
17	19	m	8	385
18	20	m	5	442
19	22	m	3	200
20	35	m	5	
21	21	m	3	400
22	22	m	6	450
23	19	m	8	520
24	20	w	3	400
25	22	m	3	
26	22	m	1	510
27	21	m	5	450
28	21	m	6	430
29	26	m	6	1400
30	22	w	4	600
31	19	w	5	660

32	21	m	6	730
33	17	m	6	670
34	21	m	5	680
35	23	m	3	510
36	26	m	6	720
37	19	m	6	600
38	23	m	3	390
39	23	m	6	370
40	22	w	6	
41	21	m	3	470
42	21	m	6	500
43	25	w	6	610
44	22	m	8	
45	22	w	6	645
46	21	m	5	550
47	21	m	3	630
48	18	w	6	340
49	23	w	6	300
50	19	m	6	375
51	18	m	6	
52	21	m	3	570
53	23	m	6	330
54	33	m	3	1350
55	21	m	3	510
56	20	w	1	182
57	23	w	2	405
58	23	w	8	490
59	21	m	3	360
60	23	m	3	430
61	21	w	2	600
62	27	w	8	515
63	21	m	6	530
64	20	m	6	341
65	22	w	3	550
66	23	m	2	525
67	22	m	3	800
68	24	m	4	550
69	21	m	8	304
70	20	m	4	460
71	20	m	1	810
72	24	m	3	500
73	19	w	1	585
74	19	m	6	440
75	20	m	3	480

76	22	w	5	960
77	18	m	8	180
78	23	m	6	410
79	22	m	5	
80	21	m	7	440
81	20	w	6	710
82	20	m	5	420
83	21	m	6	600
84	25	m	4	600
85	18	m	8	542
86	20	w	6	370
87	21	w	2	500
88	20	m	8	420
89	20	m	6	700
90	19	m	3	580
91	22	w	5	2300
92	20	w	8	570
93	18	m	6	
94	19	m	6	
95	18	m	3	1150
96	23	m	3	420
97	28	m	3	460
98	21	w	4	306
99	29	m	2	480
100	19	m	6	490

14.2 Anhang 2: Wichtige Formeln

$h = \frac{H}{n}$ $\frac{H \cdot 100}{n}$ %

$$\acute{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

Erklärung: *n*: Anzahl der Werte x_i x_i: *i*-ter Beobachtungswert)

$$Q_p = x_{[n \cdot p]+1} - f\ddot{u}r(n \cdot p) nichtganzzahlig$$

Erklärung: Die Gauss-Klammer [ ] ergibt den ganzzahligen Anteil der Zahl. Beispiel: [8,265]=8

$$Q_p = \frac{1}{2}\left(x_{n \cdot p} + x_{n \cdot p+1}\right) - f\ddot{u}r(n \cdot p) ganzzahlig$$

Erklärung: *p* ist der Prozentsatz der Werte, die unterhalb des Quantils liegen sollen.

r = (x_{max}-x_{min})

iqr = (x_{75}-x_{25})

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \acute{x})^2$$

$$\sigma = \sqrt{s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \acute{x})^2}$$

$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{x}(x_i - \acute{x}) \cdot (y_i - \acute{y})}{\sigma_x \cdot \sigma_y}$$